Lab 4b Practice: Sampling Distributions and Confidence Intervals (Mean)

Overview

In this lab we explore sampling distributions by obtaining random samples from a population. In this lab, **we consider the data set provided to be the entire population of interest** so that we may explore what happens when we repeatedly sample from the given population.

The Data: Youth Risk Behavior Surveillance System

The Youth Risk Behavior Surveillance System (YRBSS) has been conducted every two years since 1991 by the Centers for Disease Control and Prevention (CDC) in order to obtain information from adolescents regarding trends in risky behavior, such a smoking, drinking, drug use, diet, and physical activity. In 2013, 47 states participated in this school-based survey, yielding 13,583 respondents and 213 variables. Full survey and data documentation can be accessed on the CDC website. A subset of this data set which has no missing data for 16 selected variables is provided in the file `yrbss2013.csv`[1].

`age`	*Q1: How old are you?*
`gender`	*Q2: What is your sex?*
`height_m`	calculated variable: height in meters
`weight_kg`	calculated variable: weight in kilograms
`bmi`	calculated variable: body mass index=`height_m`/(`weight_kg`)^2
`BMIPCT`	calculated variable: BMI percentile for age and sex
`seatbelt`	*Q9: How often do you wear a seat belt when riding in a car driven by someone else?*
`seatbelt2`	calculated variable: `seatbelt` never vs otherwise
`ride_drunkdriver`	*Q10: During the past 30 days, have you ridden in a car or other vehicle driven by someone who had been drinking alcohol?*
`drive_drunk`	*Q11: During the past 30 days, how many times did you drive a car or other vehicle when you had been drinking alcohol?*
`drive_text`	*Q12: During the past 30 days, on how many days did you text or e-mail while driving a car or other vehicle?*
`carried_weapon`	*Q13: During the past 30 days, did you carry a weapon such as a gun, knife, or club?*
`unsafe_school`	*Q16: During the past 30 days, did you not go to school because you felt you would be unsafe at school or on your way to or from school?*
`bullied`	*Q24: During the past 12 months, have you ever been bullied on school property?*
`sad`	*Q26: During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?*
`days_smoke`	*Q33: During the past 30 days, on how many days did you smoke cigarettes?*
`days_drink`	*Q43: During the past 30 days, on how many days did you have at least one drink of alcohol?*

Practice

1. Open the `R Reference Guide` from the `Lab Content` area of PolyLearn.
2. Open `RStudio` and then open a brand new `Rmarkdown` document by clicking on the green plus sign on the top left of RStudio. Delete everything in the Rmarkdown document. Identify the `Rmarkdown` tab of your `R Reference Guide`. Copy and paste the three code chunks (header, set up chunk, and import data code chunk) into your Rmarkdown document.

[1] The variables `days_smoke` and `sad` were originally coded in categories of '0 days', '1 or 2 days', '3 to 5 days', '6 to 9 days', '10 to 19 days', '20 to 29 days', and 'All 30 days'. The number of days provided in this data set was randomly generated according to the category specified.

3. Identify the `Lab Data Sets` folder on PolyLearn and download the `yrbss2013` data set to a location on your computer (i.e., desktop, STAT 217 folder). Follow the steps in the `Importing` tab of the `R Reference Guide` to import the `yrbss2013` data set and save your import code in an R chunk.

4. Consider this data set to be a population. Write code in an R chunk to examine the population distribution of `weight_kg`. Describe the shape, center, and spread of the population distribution of weight. *(Hint: you may need to refer to R commands learned in previous labs.)*

   - shape
   - mean
   - standard deviation

5. The distribution of many sample means from samples of the same size is called the sampling distribution of the sample mean. According to the theory for the distribution of sample means, we can assume that the shape of the distribution is approximately normal when we have a normal underlying population distribution or a sample size of at least 30. Suppose we were to take many samples of size $n = 50$ from the the `weight_kg` variable. Should the distribution of sample means weights be approximately normally distributed? Why or why not?

6. According to the theory for the distribution of sample means,

   $$\text{mean} = \mu, \text{ and standard deviation} = \frac{\sigma}{\sqrt{n}}$$

   . Describe the mean and standard deviation of the distribution of the sample mean weights for samples of size 50, and fully justify your answers with explanations or calculations.

7. On a *separate* sheet of paper, use the information from the previous two questions to *sketch* the distribution of sample means for samples of size 50. Make sure it is legible and that your x-axis is labeled. You may want to check with your instructor or TA to confirm that you sketched it correctly. Include the sketch here by: taking a picture of it, emailing it to yourself, and saving it in the same location on your computer as this lab `.Rmd` file. Lastly, use the command provided to include the image in your html file (you may need to change the file name and extension in the code below, depending on how you saved your image). *Note: This command does not go in an R chunk.*

```
![](sketch.jpg)
```

8. Provide an interpretation of the your sketch from the previous question. What is this picture showing us?

9. Suppose we had taken samples of size 100 from this population rather than samples of size 50. Comment on how that would change, if at all, the *shape*, *mean*, and *standard deviation* of the distribution of sample means for samples of size 100 (compared to samples of size 50). Explain why.

   - Shape:
   - Mean:
   - Standard deviation:

10. Consider the population mean of `weight_kg` to be unknown. Suppose you can only take ONE sample from this population, and based on this ONE sample you want to *estimate* the mean weight of all teenagers. Use the following R code to take a single sample of size $n = 50$ from this variable and view the results. Change the seed to the number of your choice, and comment on how your sample compares to the population.

```
library(mosaic)
set.seed(1234)
mysample<-sample(yrbss2013$weight_kg,size=50)
```

```
mysample
hist(mysample)
favstats(mysample)
```

11. Based on the sample data obtained in the previous question, calculate an approximate 95% confidence interval for the population mean weight using the formula

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

using $t^* = 2.02$. Use R as a calculator and do the calculation in an R chunk. *Hints: (1) Enter the lower and upper bound as two separate calculations, (2) when multiplying two things you must use an asterisk... 3(5) results in an error, 3*5 results in 15.*

12. Did your confidence interval capture the actual population mean weight of teenagers? Explain.

13. Provide an interpretation of your confidence interval (you can pretend like you don't know the actual population mean).

14. Submit this lab assignment as an html compiled from R Markdown. Make sure all names of group members who contributed to this lab assignment are on the html file.